7.0 *IN VITRO* ER BINDING TEST METHOD RELIABILITY ASSESSMENT

7.1 Introduction

The ICCVAM Submission Guidelines (ICCVAM, 1999) request information about the assessment of test method reliability[1]. This includes a rationale for selecting the substances used to evaluate intra- and inter-laboratory reproducibility, discussion of the extent to which the substances tested represent the range of possible test outcomes, and a statistical analysis of intra- and inter-laboratory reproducibility. In addition, measures of central tendency and variation for historical negative and positive control data and an assessment of the historical control variability should be conducted.

However, no formal validation studies to assess *in vitro* ER binding assay inter- and intra-laboratory reproducibility have been conducted, and the nature of the current database for these assays precludes a formal analysis. Historically, investigators have used these assays primarily to gain insight into the mechanisms of the binding of a ligand to the ER, to compare the relative binding of different ligands to ER isolated from different tissues and/or species, and to understand the process of ER-induced TA. Only relatively recently have ER studies been conducted to investigate the biological activities of putative endocrine disruptors.

Despite these limitations, a quantitative assessment of IC_{50} and RBA values was conducted to assess the interlaboratory reproducibility of each of the 14 *in vitro* ER binding assays considered in this BRD. The assessment was based on the 238 substances tested in at least two assays (**Appendix E**), and was limited to individual tests that resulted in an IC_{50} or RBA value (i.e., the substance was classified as positive).

7.2 Quantitative Assessments of Interlaboratory Reproducibility

To reduce the extent of skewness in the data prior to conducting the quantitative assessments, the two outcome variables for *in vitro* ER binding assays -- the RBA and the IC_{50} values -- were

[1] Reliability is a measure of the degree to which a test can be performed reproducibly within and among laboratories over time, where reproducibility is the variability between single test results obtained in a single laboratory (intralaboratory reproducibility) or in different laboratories (interlaboratory reproducibility) using the same protocol.

transformed using the natural log. Studies that did not result in an IC_{50} and/or RBA value were excluded from consideration. Estimates of variance were compared across substances within the same assay, across substances without regard to the assay, and across assays without regard to the substances. A comparison of variances provides insight into which assays are the most reliable (i.e., all other aspects being equal, the smaller the variance, the more reliable the assay). Given the large number of data points for modeling, the general linear models (GLM) used in this analysis are robust, although some skewness may yet exist with the data. To simplify the comparison, each literature citation was considered an independent assessment (designated here as a 'reference').

As described in **Section 6**, two-way and three-way analysis of variance models were performed with random effects to estimate the intra-class correlation of substances. A high correlation value indicates that the lnRBA or $lnIC_{50}$ values are more similar within groups than among groups, where groups can be defined by assay or by reference. Estimates of variance for each model component and intra-class correlation are presented to show which factors (substance, assay, or reference) are responsible for the greatest variation in the lnRBA and $lnIC_{50}$ values. Due to limitations in the database with regard to the number of substances tested in multiple assays and to the number of independent tests performed for a substance using the same assay, the results of these analyses must be viewed with caution.

Initially, all data representing all substances, assays, and references were considered, and unique data (i.e., substances tested only in a single assay) were excluded from subsequent analyses. Information on the distribution of lnRBA and $lnIC_{50}$ values, as a function of data points, assays, and references are provided in **Section 6.2.** Consistent with the quantitative analysis on performance, the $lnIC_{50}$ and the lnRBA values for 17β-estradiol were omitted from these analyses.

7.2.1 Measures of Intra-Class Correlation

The intra-class correlation, r_I, measures the percentage of variation in y, the outcome variable, explained by a given component or set of components. The model is y = substance + assay + reference. **Table 7-1** contains the components of variance for each variable adjusted for the

other two variables (see also **Section 6**). Interpretation of this analysis is limited to factors that impact on reliability; factors that impact on assay performance are discussed in **Section 6**. From this analysis, it appears that the lnRBA or $lnIC_{50}$ values calculated for a specific substance were generally consistent irrespective of how many times a substance was tested using the same assay.

Table 7-1 Components of Variance for Each Variable Adjusted for the Other Two Variables – Reliability Assessment

	Outcome, y (% variation)	
	lnRBA	$lnIC_{50}$
Var(substance)	8.34	8.49
Var(assay)	0.38	0.34
Var(reference)	1.40	2.01
Var(error)	1.75	2.44
Corr (y_{ijk}, $y_{ij'k'}$)	0.70	0.64
Corr (y_{ijk}, $y_{ijk'}$)*	0.73	0.67
Corr (y_{ijk}, $y_{ij'k}$)	0.82	0.79

*A high correlation was found for a substance tested in the same assay (i.e., the variation in response of a substance within an assay was similar to that observed across assays).

The high correlation suggests that little variation existed in test results for an individual substance tested multiple times. However, because the majority of repeat tests were conducted on substances that were relatively potent in terms of *in vitro* ER binding, it is not known if similar variances would be found among weakly binding substances.

7.2.2 Evaluation of Substances Tested in Nine or More *In Vitro* ER Binding Assays

In this analysis, the variances for the RBA values of 12 substances that had been tested in at least 9 of the 14 *in vitro* ER binding assays were determined. The variances and sample sizes for these 12 substances are provided in **Table 7-2,** ranked in descending order according to the median RBA value based on all positive test data. Only assays for which could be calculated are included, and most of these variances were based on three or four values only. Due to the lack of sufficient data, a corresponding analysis of IC_{50} values was not conducted.

Table 7-2 Variance of lnRBA by Substance and Assay – Reliability Assessment[a]

Substance[b] (CASRN)	Median[c] RBA	#of Obs/ # Assays	hERα[d]	hERα-FP[d]	hERβ[d]	MCF-7 cytosol[d]	MUC[d]	RUC[d]	p1*	p2**
4-Hydroxy-tamoxifen (68047-06-3)	168	18/13	0.28 (3)	1.82 (3)					0.08	0.15
DES (56-53-1)	127	38/14	0.99 (3)	0.45 (4)			0.60 (7)	3.62 (11)	0.15	0.99
Estrone (53-16-7)	45	18/13				2.40 (3)		0.98 (4)	0.73	na[e]
Estriol (50-27-1)	15.8	16/12				2.42 (4)			0.53	0.64
Zearalenone (17924-92-4)	15.0	11/9	All n≤2						0.42	na
Tamoxifen (10540-29-1)	5.0	21/14	0.44 (3)					2.95 (4)	0.02	0.10
Coumestrol (479-13-0)	3.1	15/11	0.79 (3)						0.02	0.25
HPTE (2971-36-0)	1.45	12/10						1.53 (3)	0.82	na
Genistein (446-72-0)	1.30	18/11	1.07 (4)		0.97 (3)				0.11	0. 18
Bisphenol A (80-05-7)	0.031	22/14	1.36 (3)					1.25 (5)	1.25	0.60
o,p'-DDT (789-02-6)	0.038	17/12						2.97 (5)		
Kepone (143-50-0)	0.027	11/9						1.39 (3)	0.60	na

[a]Only assays where a variance could be calculated for at least one of the 12 substances are listed. The variance for a particular assay could be calculated only if a particular substance was tested three or more times in that assay; empty cells indicate insufficient data to calculate a variance. The p values could be calculated only if there were two observations from at least three or more assays; a missing p-value indicates insufficient data.

[b]Substances that had been tested in at least nine of the 14 *in vitro* ER binding assays; DES = diethylstilbestrol; *o,p'*-DDT = *o,p'*-dichlorodiphenyltrichloroethane; HPTE = (2,2-Bis(*p*-hydroxyphenyl)-1,1,1,-trichloroethane

[c]The median RBA value across assays, based on positive test data.

[d]The numbers in parenthesis indicate the numbers of replicate tests.

[e]na = No p value could be calculated since there was either no values or only one value per assay x response combination.

*p1 tests whether there is a significant difference among all assays used; unadjusted for references.

**p2 tests whether there is a significant difference among all assays used; adjusted for references.

The similarity between the p1 and p2 values for most of these substances suggests that there were no significant differences in the performance of the assays by different laboratories (a measure of assay reliability). However, DES and coumestrol exhibited considerably less variability when the analysis was adjusted for the reference (i.e., p2 is much greater than p1), suggesting that laboratory-specific differences in testing of this substance were responsible.

7.2.3 Variability in Standard Deviation for lnRBA and $lnIC_{50}$ Values by *In Vitro* ER Binding Assay

Because of insufficient data on substances tested in the same assay within or across laboratories, separate correlations between pairs of the 14 assays were not calculated. However, standard deviations of the mean of the lnRBA and the $lnIC_{50}$ values were inspected to see which assays have the least or the most variability in their responses (**Table 7-3**). A major limitation of this analysis is that the same substances were not tested in each assay. An additional limitation is the varied number of substances tested more than once in each assay. These limitations will affect any interpretation of the results. The assays in **Table 7-3** are sorted in descending order based on the number of different substances tested in each assay.

Not unexpectedly, for the same set of substances, there appears to be more variability in the IC_{50} values than in the corresponding normalized RBA values. The standard deviations for the majority of substances clustered around a median of 3.23 for the lnRBA values and 3.52 for the $lnIC_{50}$ values. The standard deviations for the lnRBA values vary from a low of 2.92 for the rERβ assay to a high of 5.09 for the RBC assay, while the corresponding standard deviations for the $lnIC_{50}$ values vary from a low of 2.95 for the GST-hERα assay to a high of 4.85 for the RBC assay. The MCF-7 cells, hERβ, GST-hERα and MCF-7 cytosol assays exhibited similar and relatively lower standard deviations for lnRBA values, while the GST-aERdef, hERβ and GST-rtERdef assays exhibited similar and relatively lower standard deviations for $lnIC_{50}$ values. The RBC assay exhibited the largest standard deviation for both the lnRBA and $lnIC_{50}$ values. Based on this analysis, the hERβ assay appears to be the most reliable, while the RBC assay appears to be the least reliable. However, these conclusions must take into account the number of substances that have been tested in each ER assay and, although not specified, the number of

laboratories that generated the data. In general, the standard deviation increases as the number of substances tested in an assay increases or as more laboratories are involved.

Table 7-3 Standard Deviation for lnRBA and $lnIC_{50}$ Values for *In Vitro* ER Binding Assays

Assay	Number of Different Substances	LnRBA		$lnIC_{50}$	
		Standard Deviation	n[a]	Standard Deviation	n[a]
RUC	100	4.34	164	4.30	90
hERα	87	3.26	112	3.69	24
hERβ	74	3.03	91	3.24	30
MCF-7 cytosol	63	3.07	72	3.55	15
MCF-7 cells	58	2.94	49	3.46	2
GST-rtERdef	43	3.20	43	3.26	43
MUC	33	3.49	49	3.37	35
GST-hERαdef	29	3.01	29	2.95	29
rERβ	28	2.92	28	-	0
GST-aERdef	25	3.19	25	3.15	25
GST-cERdef	21	3.68	21	3.68	21
RBC	21	5.09	22	4.85	8
GST-mERαdef	19	3.53	19	3.52	19
hERα-FP	19	3.84	28	3.80	28

[a]Total number of data points considered in the analysis.

7.2.4 Variability in the IC_{50} for 17β-Estradiol

The most extensive database within and across assays is for 17β-estradiol, the natural estrogen commonly used as the reference substance in *in vitro* ER binding assays for calculating the RBA value of a test substance. However, because the RBA value for this substance is arbitrarily set at 100, this measure of binding cannot be analyzed for variability. In contrast, an analysis of the IC_{50} values of 17β-estradiol, where reported, provides a means for assessing assay reproducibility. Fifty-eight IC_{50} values were available for 17β-estradiol in the ER binding database (**Appendix D**). The variability in the natural log of IC_{50} values of 17β-estradiol was compared across assays. As the sample size within each assay is quite small, only descriptive statistics of this parameter are presented (**Table 7-4**). The IC_{50} values are sorted in descending order based on the number of times 17β-estradiol was tested in each assay.

Table 7-4 Standard Deviation for IC_{50} Values Obtained for 17β-Estradiol

Assay	N[a]	Standard Deviation[b]
RUC	13	0.90
MUC	10	2.21
hERα	9	0.99
hERα-FP	7	0.61
hERβ	4	0.78
MCF-7 cytosol	3	4.06
GST-hERαdef	3	0.44
GST-rtERdef	2	0.044
GST-aERdef	2	0.15
GST-cERdef	1	
GST-mERαdef	1	
MCF-7 cells	1	
RBC	1	
rERβ	1	

[a]Number of data points considered.
[b]Standard deviations could not be calculated for single test data.

In this analysis, the greater the standard deviation, the less reliable is the assay. Since the IC_{50} values of 17β-estradiol for the rERβ, RBC, MCF-7 cells, GST-mERαdef, and GST-cERdef assays were reported by one laboratory only, no standard deviations could be calculated. Although the standard deviations for the IC_{50} values were very small for the GST-aERdef, GST-rtERdef, and GST-hERαdef assays, only two or three data points were reported for these assays. Among the assays with at least six data points (hERα-FP, MUC, RUC, hERα), the standard deviations in the $lnIC_{50}$ values are generally similar except for the hERα-FP assay where the value is smaller. Although this decreased standard deviation suggests that the hERα-FP assay is the most reliable of these four assays, the hERα-FP assay data were generated by fewer laboratories, which may have impacted on the extent of variability.

7.3 Reliability of *In Vitro* ER Binding Assays

The *in vitro* ER binding assays that are the most useful as a screen for endocrine disruptors are those that are the most sensitive (i.e., have the greatest ability to detect weak ER-binding substances) (see **Section 6**) and the most reliable (i.e., exhibit the lowest variance). The results

of the quantitative assessments of the comparative reliability of the 14 *in vitro* ER binding assays evaluated in this BRD are summarized in **Table 7-5**.

Table 7-5 Summary of *In Vitro* ER Binding Assay Reliability

Assay	lnRBA[a]	$lnIC_{50}$[a]	IC_{50} 17β-Estradiol[c]
RUC	-	-	0
GST-aERdef	0	+	?
GST-cERdef	-	0	?
GST-hERαdef	+	+	?
GST-mERαdef	0	0	?
GST-rtERdef	0	+	?
hERα	0	0	0
hERα-FP	-	-	+
hERβ	+	+	?
MCF-7 cells	+	0	?
MCF-7 cytosol	+	0	?
MUC	0	+	0
RBC	-	-	?
rERβ	+	?	?

[a]Reliability based on standard error term for lnRBA values (**Table 7-3**); more reliable = +; average reliability = 0; less reliable = -.

[a]Reliability based on standard error term for $lnIC_{50}$ values (**Table 7-3**); more reliable = +; average reliability = 0; less reliable = -; ? = the number of observations was too small to make a determination.

[c]Reliability based on variance analysis of $lnIC_{50}$ values for 17β-estradiol (**Table 7-4**); most reliable = +, average reliability = 0; least reliable = -; ? = the number of observations was too small to make a useful determination.

Based on a weight-of-evidence approach, the GST-hERαdef and hERβ assays appear to offer the greatest overall reliability (both assays had two reliable categories). However, due to the absence of formal validation studies to assess reliability and to the paucity of the data on which this reliability assessment is made, the decision to select any one assay or group of assays over another appears to be arbitrary.

7.4 Conclusions and Recommendations

Although a large number of substances have been tested in *in vitro* ER binding assays, relatively few substances have been tested more than once in the same assay or in multiple assays, and no

formal validation studies have been conducted to assess reliability. A quantitative assessment was conducted using the available IC_{50} and RBA data after being log-normal transformed to reduce possible skewness. One limitation of this approach was that situations in which a substance was classified as negative and positive in different tests using the same assay was not considered.

An analysis of the variances for the RBA values of 12 substances that had been tested in at least nine of the 14 *in vitro* ER binding assays suggested that there were no significant differences in the reliability of the assays as performed by different laboratories. Inspection of the standard errors of the mean of the lnRBA and the $lnIC_{50}$ values suggested that the RUC assay appeared to be the most reliable, while the RBC assay appeared to be the least reliable. A major limitation of this analysis is that the same substances were not tested in each assay and that the number of substances that have been tested in each ER assay or the number of laboratories that generated the data was not considered.

A comparison of the variability in lnRBA and $lnIC_{50}$ values across assays, ignoring substance effects, indicated that the GST-hERαdef and hERβ assays were the most consistent and the RBC assay was the least consistent among the 14 assays evaluated. An analysis of the variability in the $lnIC_{50}$ for 17β-estradiol, the reference estrogen for these assays, indicated that the most consistent results were obtained with the hERα-FP assay, while the MUC, RUC, and hERα assays exhibited somewhat greater but comparable variances. The low variability associated with the hERα-FP assay, however, might be a reflection of the small number of laboratories that have reported RBA values using this method. Data were too limited to evaluate the other *in vitro* ER binding assays.

Taking into account the available *in vitro* ER binding assay database and the various quantitative assessments conducted on the 14 *in vitro* ER binding assays considered in this BRD, the following recommendation can be made in regard to the use of such assays as screening test methods within a battery of Tier 1 endocrine disruptor tests.

- Despite inferences that the GST-hERαdef and hERβ assays appear to be the most reliable among the 14 *in vitro* ER binding assays considered in this BRD, an adequate assessment of

assay reliability cannot be performed based on the limited database available. However, it might be expected that assays that use semi-purified or purified ER proteins would be more reliable than those based on extracts of ER from animal tissues.

- It is essential that validation studies be conducted to assess assay reliability and that these validation studies use appropriate substances covering the range of expected RBA values. A list of potential test substances for use in such a validation effort is provided in **Section 12**.